Available online at www.sciencedirect.com
ScienceDirect
Theoretical Computer Science 387 (2007) 4–17
Theoretical Computer Science
www.elsevier.com/locate/tcs

Stochastic complexity for mixture of exponential families in generalized variational Bayes

Kazuho Watanabe[a,*], Sumio Watanabe[b]

[a] *Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology, Japan*
[b] *P&I Lab., Tokyo Institute of Technology, Japan*

Abstract

The Variational Bayesian learning, proposed as an approximation of the Bayesian learning, has provided computational tractability and good generalization performance in many applications. However, little has been done to investigate its theoretical properties.

In this paper, we discuss the Variational Bayesian learning of the mixture of exponential families and derive the asymptotic form of the stochastic complexities in a generalized setting of the prior distribution. We show that the stochastic complexities become smaller than those of regular statistical models, which implies that the advantage of the Bayesian learning still remains in the Variational Bayesian learning. Stochastic complexity, which is called the marginal likelihood or the free energy, not only becomes important in addressing the model selection problem but also enables us to discuss the accuracy of the Variational Bayesian approach as an approximation of the true Bayesian learning. The main result also shows the effects of the prior distribution under the generalized setting.

Keywords: Mixture model; Exponential family; Variational Bayes; Generalized Bayes; Stochastic complexity; Free energy; Kullback information; Non-regular model

1. Introduction

The Bayesian learning is one of the most powerful methods for models with hidden variables. The Variational Bayesian (VB) framework was proposed as an approximation of the Bayesian learning [3,7]. This framework provides an effective iterative algorithm to compute the posterior distributions over the hidden variables and parameters. The Variational Bayesian learning has been applied to various learning machines and it has performed good generalization with only modest computational costs compared to Markov chain Monte Carlo (MCMC) methods that are the major schemes of the Bayesian learning.

In spite of its tractability and its wide range of applications, little has been done to investigate the theoretical properties of the Variational Bayesian learning itself. Although the Variational Bayesian framework is an

* Corresponding address: Department of Complexity Science and Engineering, The University of Tokyo, Tokyo, Japan.
E-mail addresses: kazuho@mns.k.u-tokyo.ac.jp (K. Watanabe), swatanab@pi.titech.ac.jp (S. Watanabe).

0304-3975/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2007.07.039

approximation, questions like how accurately it approximates the true one remained unanswered until quite recently. To address these issues, the asymptotic form of the stochastic complexity in the Variational Bayesian learning was clarified and the accuracy of the Variational Bayesian learning was discussed in the case of mixtures of Gaussians [14] and mixtures of exponential families [15].

In this paper, we focus on the generalized Variational Bayesian learning of the mixtures of exponential families which include mixtures of distributions such as Gaussian, binomial and gamma. In this generalized framework, the prior distribution is controlled by sequences α_n and β_n depending on the sample size n while in the original Bayesian or Variational Bayesian learning, it is fixed with respect to n. We consider the case in which the true distribution is contained in the learner model. In this case, the parameters are non-identifiable in mixture models due to their hidden variables. Hence, mixture models are known to be non-regular statistical models. In some recent studies, the Bayesian stochastic complexities of non-regular models have been clarified and it has been proven that they become smaller than those of regular models [16–18]. This indicates an advantage of the Bayesian learning which is typical in non-regular models. Therefore, analyzing the stochastic complexity in this case is most valuable for comparing the Variational Bayesian learning with the true Bayesian learning. Furthermore, this analysis is necessary and essential for addressing the model selection and hypothesis testing problems.

As the main result, we derive the upper and lower bounds of the stochastic complexity in the generalized Variational Bayesian learning of the mixture of exponential families and show that the stochastic complexity becomes smaller than those of regular models. Since the derived bounds show us the accuracy of the Variational Bayesian learning as an approximation method, our result implies that the advantage of the Bayesian learning still remains in the Variational Bayesian learning. In addition, the derived bounds give us an indication of how the prior distribution influences the process of the learning. The effects of the sequences α_n, β_n and that of the prior hyperparameter are discussed.

The paper is organized as follows. In Section 2, we introduce the mixture of exponential family model. In Section 3, we describe the generalized Bayesian learning. In Section 4, the Variational Bayesian framework is described and the variational posterior distribution for the mixture of exponential family model is derived. In Section 5, we present our main result (Theorem 2). The main theorem is proved in Appendix. Discussion and conclusion follow in Sections 6 and 7.

2. Mixture of exponential family

Denote by $c(x|b)$ a probability density function of the input $x \in R^N$ given an M-dimensional parameter vector $b = (b^{(1)}, b^{(2)}, \ldots, b^{(M)})^{\mathrm{T}} \in B$ where B is a subset of R^M. The general mixture model $p(x|\theta)$ with a parameter vector θ is defined by

$$p(x|\theta) = \sum_{k=1}^{K} a_k c(x|b_k),$$

where K is the number of components and $\{a_k | a_k \geq 0, \sum_{k=1}^{K} a_k = 1\}$ is the set of mixing proportions. The parameter θ of the model is $\theta = \{a_k, b_k\}_{k=1}^{K}$.

A model $p(x|\theta)$ is called a mixture of exponential family (MEF) model or exponential family mixture model if the probability distribution $c(x|b)$ is given by the following form,

$$c(x|b) = \exp\{b \cdot f(x) + f_0(x) - g(b)\}, \tag{1}$$

where $b \in B$ is called the natural parameter, $b \cdot f(x)$ is the inner product with the vector $f(x) = (f_1(x), \ldots, f_M(x))^{\mathrm{T}}$, $f_0(x)$ and $g(b)$ are real-valued functions of the input x and the parameter b, respectively [5]. Suppose functions $f_1, \ldots, f_M$ and a constant function are linearly independent and the effective number of parameters in a single component distribution $c(x|b)$ is M.

The conjugate prior distribution $\varphi(\theta)$ for the mixture of exponential family model is defined by the product of the following two distributions on $\mathbf{a} = \{a_k\}_{k=1}^{K}$ and $\mathbf{b} = \{b_k\}_{k=1}^{K}$

$$\varphi(\mathbf{a}) = \frac{\Gamma(K\phi_0)}{\Gamma(\phi_0)^K} \prod_{k=1}^{K} a_k^{\phi_0 - 1}, \tag{2}$$

$$\varphi(\mathbf{b}) = \prod_{k=1}^{K} \varphi(b_k) = \prod_{k=1}^{K} \frac{1}{C(\xi_0, \nu_0)} \exp\{\xi_0(b_k \cdot \nu_0 - g(b_k))\}, \tag{3}$$

where the function $C(\xi, \mu)$ of $\xi \in R$ and $\mu \in R^M$ is defined by

$$C(\xi, \mu) = \int \exp\{\xi(\mu \cdot b - g(b))\} \mathrm{d}b. \tag{4}$$

Here $\xi_0 > 0$, $\nu_0 \in R^M$ and $\phi_0 > 0$ are constants called hyperparameters.

In the generalized Bayesian framework, the prior distribution $\varphi(\theta)$ is replaced by

$$\varphi_n(\theta) = \varphi_n(\mathbf{a})\varphi_n(\mathbf{b}), \tag{5}$$

where

$$\varphi_n(\mathbf{a}) = \frac{1}{C_{\alpha_n}} \varphi(\mathbf{a})^{\alpha_n}, \tag{6}$$

$$\varphi_n(\mathbf{b}) = \prod_{k=1}^{K} \varphi_n(b_k) = \frac{1}{C_{\beta_n}} \varphi(\mathbf{b})^{\beta_n}. \tag{7}$$

Here $\varphi_n(b_k) = 1/C_{\beta_n}^{1/K} \varphi(b_k)^{\beta_n}$ and C_{α_n}, C_{β_n} are the normalization constants. $\alpha_n \geq 1$, $\beta_n \geq 1$ are monotone non-decreasing sequences depending on the sample size n. If $\alpha_n = \beta_n = 1$, this reduces to the original Bayesian framework.

The mixture model can be rewritten as follows by using a hidden variable $y = (y^1, \ldots, y^K) \in \{(1, 0, \ldots, 0), (0, 1, \ldots, 0), \ldots, (0, 0, \ldots, 1)\}$,

$$p(x, y|\theta) = \prod_{k=1}^{K} \Big[a_k c(x|b_k) \Big]^{y^k}.$$

The hidden variable y is not observed and is representing a component from which the datum x is generated. If and only if the datum x is from the kth component, then $y^k = 1$.

The mixture model is a non-regular statistical model, since the parameters are non-identifiable. More specifically, if the true distribution is realized by a model with the smaller number of components, the true parameter is not a point but an analytic set with singularities. If the model parameters are non-identifiable, the usual asymptotic theory of regular statistical models cannot be applied. Some studies have revealed that the mixture model has quite different properties from those of regular statistical models [8,17].

3. The generalized bayesian learning

Suppose n training samples $X^n = \{x_1, \ldots, x_n\}$ are independently and identically taken from the true distribution $p_0(x)$. In the Bayesian learning of a model $p(x|\theta)$ whose parameter is θ, first, the prior distribution $\varphi(\theta)$ on the parameter θ is set. In the generalized Bayesian framework, $\varphi_n(\theta)$ is defined by Eq. (5). Then the posterior distribution $p(\theta|X^n)$ is computed from the dataset and the prior $\varphi_n(\theta)$ by

$$p(\theta|X^n) = \frac{1}{Z(X^n)} \varphi_n(\theta) \prod_{i=1}^{n} p(x_i|\theta),$$

where $Z(X^n)$ is the normalization constant that is also known as the marginal likelihood or the evidence of the dataset X^n [10]. In this paper, as the monotone non-decreasing sequence $\alpha_n \geq 1$ in Eq. (6), we use the sequence that is smaller order than $\log n$ and consider these two types, α_n is bounded or $\alpha_n \to \infty$ as n tends to infinity. As the monotone non-decreasing sequence β_n, the same two types are considered.

The Bayesian predictive distribution $p(x|X^n)$ is given by averaging the model over the posterior distribution as follows,

$$p(x|X^n) = \int p(x|\theta)p(\theta|X^n)\mathrm{d}\theta. \tag{8}$$

The stochastic complexity $F(X^n)$ is defined by

$$F(X^n) = -\log Z(X^n), \tag{9}$$

which is also called the free energy and is important in most data modelling problems. Practically, it is used as a criterion by which the learner model is selected and the hyperparameters in the prior are optimized [1,13].

The Bayesian posterior can be rewritten as

$$p(\theta|X^n) = \frac{1}{Z_0(X^n)} \exp(-nH_n(\theta))\varphi_n(\theta), \tag{10}$$

where $H_n(\theta)$ is the empirical Kullback information,

$$H_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p_0(x_i)}{p(x_i|\theta)}, \tag{11}$$

and $Z_0(X^n)$ is the normalization constant. Putting $S(X^n) = -\sum_{i=1}^{n} \log p_0(x_i)$, we define the normalized stochastic complexity $F_0(X^n)$ by

$$F_0(X^n) = F(X^n) - S(X^n). \tag{12}$$

It is noted that the empirical entropy $S(X^n)$ does not depend on the model $p(x|\theta)$ and its expectation value over all sets of training samples is equal to nS where $S = -\int p_0(x)\log p_0(x)\mathrm{d}x$ is the entropy. Therefore minimization of $F(X^n)$ is equivalent to that of $F_0(X^n)$.

We define the average normalized stochastic complexity $F_0(n)$ by

$$F_0(n) = E_{X^n}[F_0(X^n)], \tag{13}$$

where $E_{X^n}[\cdot]$ denotes the expectation value over all sets of training samples.

Recently, it was proved that, in the case of the conventional Bayes ($\alpha_n = \beta_n = 1$), the average normalized stochastic complexity $F_0(n)$ has the following asymptotic form [16],

$$F_0(n) \simeq \lambda \log n - (m-1)\log\log n + O(1), \tag{14}$$

where λ and m are the rational number and the natural number respectively which are determined by the singularities of the set of true parameters. In regular statistical models, 2λ is equal to the number of parameters and $m = 1$, whereas in non-regular models such as the mixture model, 2λ is not larger than the number of parameters and $m \geq 1$. This means non-regular models have an advantage in the Bayesian learning because the stochastic complexity corresponds to the cumulative loss of the Bayesian predictive distribution and the redundancy of the Bayesian method in coding [6].

However, in order to carry out the Bayesian learning practically, one computes the stochastic complexity or the predictive distribution by integrating over the posterior distribution, which typically cannot be performed analytically. As an approximation, the Variational Bayesian framework was proposed [3,4,7].

4. The variational Bayesian learning

4.1. The variational Bayesian framework

Using the likelihood on the complete data $\{X^n, Y^n\}$ added the corresponding hidden variables $Y^n = \{y_1, \ldots, y_n\}$, we can rewrite the stochastic complexity Eq. (9) as

$$F(X^n) = -\log \int \sum_{Y^n} p(X^n, Y^n, \theta)\mathrm{d}\theta,$$

where $p(X^n, Y^n, \theta) = \varphi_n(\theta) \prod_{i=1}^n p(x_i, y_i|\theta)$ and the sum over Y^n ranges over all possible values of all hidden variables.

In the Variational Bayesian framework, the Bayesian posterior distribution $p(Y^n, \theta|X^n)$ of the hidden variables and the parameters is approximated by the variational posterior distribution $q(Y^n, \theta|X^n)$, which factorizes as

$$q(Y^n, \theta|X^n) = Q(Y^n|X^n)r(\theta|X^n), \tag{15}$$

where $Q(Y^n|X^n)$ and $r(\theta|X^n)$ are probability distributions on the hidden variables and the parameters respectively. The variational posterior $q(Y^n, \theta|X^n)$ is chosen so that it minimizes the functional $\overline{F}[q]$ defined by

$$\overline{F}[q] = \sum_{Y^n} \int q(Y^n, \theta|X^n) \log \frac{q(Y^n, \theta|X^n)}{p(X^n, Y^n, \theta)} \mathrm{d}\theta, \tag{16}$$

$$= F(X^n) + K(q(Y^n, \theta|X^n)||p(Y^n, \theta|X^n)), \tag{17}$$

where $K(q(Y^n, \theta|X^n)||p(Y^n, \theta|X^n))$ is the Kullback information between the true Bayesian posterior $p(Y^n, \theta|X^n)$ and the variational posterior $q(Y^n, \theta|X^n)$.[1] This leads to the following theorem. The proof is well-known [4,12].

Theorem 1. *If the functional $\overline{F}[q]$ is minimized under the constraint Eq. (15) then the variational posteriors, $r(\theta|X^n)$ and $Q(Y^n|X^n)$, satisfy*

$$r(\theta|X^n) = \frac{1}{C_r} \varphi_n(\theta) \exp \left\langle \log p(X^n, Y^n|\theta) \right\rangle_{Q(Y^n|X^n)}, \tag{18}$$

and

$$Q(Y^n|X^n) = \frac{1}{C_Q} \exp \left\langle \log p(X^n, Y^n|\theta) \right\rangle_{r(\theta|X^n)}, \tag{19}$$

where C_r and C_Q are the normalization constants.[2]

Hereafter, we omit the condition X^n of the variational posteriors and abbreviate them to $q(Y^n, \theta)$, $Q(Y^n)$ and $r(\theta)$.

Note that Eqs. (18) and (19) give only a necessary condition that $r(\theta)$ and $Q(Y^n)$ minimize the functional $\overline{F}[q]$. The variational posteriors that satisfy Eqs. (18) and (19) are searched by an iterative algorithm.

We define the stochastic complexity in the Variational Bayesian learning $\overline{F}(X^n)$ by the minimum value of the functional $\overline{F}[q]$ attained by the above optimal variational posteriors, that is,

$$\overline{F}(X^n) = \min_{r,Q} \overline{F}[q].$$

Since $\overline{F}(X^n)$, the stochastic complexity in the Variational Bayesian learning, gives the upper bound of the true stochastic complexity $F(X^n)$, $\overline{F}(X^n)$ itself is an estimate of $F(X^n)$ and is used for the model selection in the Variational Bayesian learning [4]. Moreover, from Eq. (17), it is noted that the difference between $\overline{F}(X^n)$ and the original stochastic complexity $F(X^n)$ is the Kullback information from the variational posterior to the true posterior, which shows us the accuracy of the Variational Bayesian approach as an approximation of the true Bayesian learning.

We define the normalized stochastic complexity $\overline{F}_0(X^n)$ in the Variational Bayesian learning by

$$\overline{F}_0(X^n) = \overline{F}(X^n) - S(X^n). \tag{20}$$

Putting Eq. (19) into Eq. (16) gives the following lemma.

Lemma 1.

$$\overline{F}_0(X^n) = \min_{r(\theta)} \{K(r(\theta)||\varphi_n(\theta)) - (\log C_Q + S(X^n))\}, \tag{21}$$

where $C_Q = \sum_{Y^n} \exp \left\langle \log p(X^n, Y^n|\theta) \right\rangle_{r(\theta)}$.

[1] Throughout this paper, we use the notation $K(q(x)||p(x))$ for the Kullback information from a distribution $q(x)$ to a distribution $p(x)$, that is,

$$K(q(x)||p(x)) = \int q(x) \log \frac{q(x)}{p(x)} \mathrm{d}x.$$

[2] For an arbitrary distribution $p(x)$, $\langle \cdot \rangle_{p(x)}$ denotes the expectation over $p(x)$.

4.2. Variational posterior for mixture of exponential family model

In this subsection, we derive the variational posterior $r(\theta)$ for the mixture of exponential family model based on Eq. (18) and then define the variational parameter and the variational estimator for this model.

Using the complete data $\{X^n, Y^n\} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, we put

$$\overline{y}_i^k = \langle y_i^k \rangle_{Q(Y^n)}, \quad n_k = \sum_{i=1}^n \overline{y}_i^k, \quad \text{and} \quad \nu_k = \frac{1}{n_k} \sum_{i=1}^n \overline{y}_i^k f(x_i),$$

where $y_i^k = 1$ if the ith datum x_i is from the kth component, if otherwise, $y_i^k = 0$. The variable n_k is the expected number of the data that are estimated to be from the kth component. Note that the variables n_k and ν_k satisfy the constraints $\sum_{k=1}^K n_k = n$ and $\sum_{k=1}^K n_k \nu_k = \sum_{i=1}^n f(x_i)$. From Eq. (18) and the respective prior Eqs. (2) and (3), the variational posterior $r(\theta) = r(\mathbf{a})r(\mathbf{b})$ is obtained as the product of the following two distributions,

$$r(\mathbf{a}) = \frac{\Gamma(n + K\phi_n)}{\prod_{k=1}^K \Gamma(n_k + \phi_n)} \prod_{k=1}^K a_k^{n_k + \phi_n - 1}, \tag{22}$$

$$r(\mathbf{b}) = \prod_{k=1}^K r(b_k) = \prod_{k=1}^K \frac{1}{C(\gamma_k, \overline{\mu}_k)} \exp\{\gamma_k(\overline{\mu}_k \cdot b_k - g(b_k))\}, \tag{23}$$

where $\phi_n = \alpha_n(\phi_0 - 1) + 1$, $\xi_n = \xi_0 \beta_n$, $\overline{\mu}_k = \frac{n_k \nu_k + \xi_n \nu_0}{n_k + \xi_n}$ and $\gamma_k = n_k + \xi_n$.

Let

$$\overline{a}_k = \langle a_k \rangle_{r(\mathbf{a})} = \frac{n_k + \phi_n}{n + K\phi_n} \qquad \left(\frac{\phi_n}{n + K\phi_n} \le \overline{a}_k \le 1 - \frac{(K-1)\phi_n}{n + K\phi_n} \right), \tag{24}$$

$$\overline{b}_k = \langle b_k \rangle_{r(b_k)} = \frac{1}{\gamma_k} \frac{\partial \log C(\gamma_k, \overline{\mu}_k)}{\partial \overline{\mu}_k}, \tag{25}$$

and define the variational parameter $\overline{\theta}$ by

$$\overline{\theta} = \langle \theta \rangle_{r(\theta)} = \{\overline{a}_k, \overline{b}_k\}_{k=1}^K. \tag{26}$$

It is noted that $\overline{b}_k$ is the expectation parameter of b_k with the variational posterior $r(b_k)$. It is also noted that the variational posterior $r(\theta)$ and C_Q in Eq. (19) are parameterized by the variational parameter $\overline{\theta}$. Therefore, we denote them as $r(\theta|\overline{\theta})$ and $C_Q(\overline{\theta})$ henceforth. We define the variational estimator $\overline{\theta}_{vb}$ by the variational parameter $\overline{\theta}$ that attains the minimum value of the normalized stochastic complexity $\overline{F}_0(X^n)$. Then, Lemma 1 claims that

$$\overline{\theta}_{vb} = \operatorname*{argmin}_{\overline{\theta}} \{K(r(\theta|\overline{\theta})||\varphi_n(\theta)) - (\log C_Q(\overline{\theta}) + S(X^n))\}. \tag{27}$$

In the Variational Bayesian learning, the variational parameter $\overline{\theta}$ is updated iteratively to find the optimal solution $\overline{\theta}_{vb}$. Therefore, our aim is to evaluate the minimum value of the right-hand side of Eq. (27) as a function of the variational parameter $\overline{\theta}$.

5. Main results

The average normalized stochastic complexity $\overline{F}_0(n)$ in the Variational Bayesian learning is defined by

$$\overline{F}_0(n) = E_{X^n}[\overline{F}_0(X^n)]. \tag{28}$$

We assume the following conditions.

(i) The true distribution $p_0(x)$ is represented by a mixture of exponential family model $p(x|\theta_0)$ which has K_0 components and the parameter $\theta_0 = \{a_k^*, b_k^*\}_{k=1}^{K_0}$,

$$p(x|\theta_0) = \sum_{k=1}^{K_0} a_k^* \exp\{b_k^* \cdot f(x) + f_0(x) - g(b_k^*)\},$$

where $b_k^* \in R^M$ and $b_k^* \neq b_j^* (k \neq j)$. And suppose that the true distribution can be realized by the model, that is, the model $p(x|\theta)$ has K components,

$$p(x|\theta) = \sum_{k=1}^{K} a_k \exp\{b_k \cdot f(x) + f_0(x) - g(b_k)\},$$

and $K \geq K_0$ holds.

(ii) The prior distribution of the parameters is the conjugate prior $\varphi_n(\theta) = \varphi_n(\mathbf{a})\varphi_n(\mathbf{b})$ where $\varphi_n(\mathbf{a})$ and $\varphi_n(\mathbf{b})$ are given by Eqs. (6) and (7). Also, $\varphi(\mathbf{b})$ is bounded.

(iii) Regarding the distribution $c(x|b)$ of each component, the Fisher information matrix

$$I(b) = \frac{\partial^2 g(b)}{\partial b \partial b}$$

satisfies $0 < |I(b)| < +\infty$, for arbitrary $b \in B$.[3] The function $\mu \cdot b - g(b)$ has a stationary point at $\hat{b}$ in the interior of B for each $\mu \in R^M$.

Under these conditions, we prove the following theorem. The proof is given in Appendix.

Theorem 2 (*Main Result*). *Assume the conditions* (i), (ii) *and* (iii). *Then the average normalized stochastic complexity* $\overline{F}_0(n)$ *defined by Eq.* (28) *satisfies followings,*

(I) *If* α_n *and* β_n *are bounded, then*

$$\underline{\lambda} \log n + E_{X^n}\left[nH_n(\overline{\theta}_{vb})\right] + C_1 \leq \overline{F}_0(n) \leq \overline{\lambda} \log n + C_2, \tag{29}$$

for an arbitrary natural number n, where C_1 *and* C_2 *are constants independent of n. Let* $\alpha^* = \lim_{n\to\infty} \alpha_n$ *and* $\phi^* = \alpha^*(\phi_0 - 1) + 1$, *the coefficients* $\underline{\lambda}$ *and* $\overline{\lambda}$ *are given by*

$$\underline{\lambda} = \begin{cases} (K-1)\phi^* + \frac{M}{2} & \left(\phi_0 \leq 1 + \frac{M-1}{2\alpha^*}\right), \\ \frac{MK+K-1}{2} & \left(\phi_0 > 1 + \frac{M-1}{2\alpha^*}\right), \end{cases} \tag{30}$$

and

$$\overline{\lambda} = \begin{cases} (K-K_0)\phi^* + \frac{MK_0+K_0-1}{2} & \left(\phi_0 \leq 1 + \frac{M-1}{2\alpha^*}\right), \\ \frac{MK+K-1}{2} & \left(\phi_0 > 1 + \frac{M-1}{2\alpha^*}\right). \end{cases} \tag{31}$$

(II) *If* $\alpha_n \to \alpha^* < \infty$, $\beta_n \to \infty$ *and* $\beta_n / \log n \to 0$ *as n tends to infinity, then*

$$\underline{\lambda} \log n + E_{X^n}\left[nH_n(\overline{\theta}_{vb})\right] + o(\log n) \leq \overline{F}_0(n) \leq \overline{\lambda} \log n + o(\log n), \tag{32}$$

where the coefficients $\underline{\lambda}$ *and* $\overline{\lambda}$ *are given by Eqs.* (30) *and* (31).

(III) *If* $\alpha_n \to \infty$, $\alpha_n / \log n \to 0$, $\beta_n / \log n \to 0$ *as n tends to infinity and* $\phi_0 > 1$, *then*

$$\lambda_{\mathrm{BIC}} \log n + E_{X^n}\left[nH_n(\overline{\theta}_{vb})\right] + o(\log n) \leq \overline{F}_0(n) \leq \lambda_{\mathrm{BIC}} \log n + o(\log n), \tag{33}$$

where

$$\lambda_{\mathrm{BIC}} = \frac{MK + K - 1}{2}. \tag{34}$$

[3] $\frac{\partial^2 g(b)}{\partial b \partial b}$ denotes the matrix whose ijth entry is $\frac{\partial^2 g(b)}{\partial b^{(i)} \partial b^{(j)}}$ and $|\cdot|$ denotes the determinant of a matrix.

This theorem shows the asymptotic form of the average stochastic complexity in the generalized Variational Bayesian learning. The coefficient $\overline{\lambda}$, $\underline{\lambda}$ and λ_{BIC} are identified by K, K_0, that are the numbers of components of the learner and the true distribution, the number of parameters M of each component and the hyperparameter ϕ_0 of the conjugate prior given by Eq. (2). The conventional Variational Bayesian learning corresponds to the case [I] when $\alpha_n = \beta_n = 1$.

In this theorem, $nH_n(\overline{\theta}_{vb})$ is equal to $-\sum_{i=1}^n \log p(x_i|\overline{\theta}_{vb}) - S(X^n)$, and the term $-\frac{1}{n}\sum_{i=1}^n \log p(x_i|\overline{\theta}_{vb})$ is a training error which is computable during the learning. If the term $E_{X^n}[nH_n(\overline{\theta}_{vb})]$ is a bounded function of n, then it immediately follows from this theorem that in the case [I] for example,

$$\underline{\lambda}\log n + O(1) \leq \overline{F}_0(n) \leq \overline{\lambda}\log n + O(1),$$

where $O(1)$ is a bounded function of n. In certain cases, such as binomial mixtures, it is actually a bounded function of n. In the case of Gaussian mixtures, if $B = R^N$, it is conjectured that the minus log likelihood ratio $\min_\theta nH_n(\theta)$, a lower bound of $nH_n(\overline{\theta}_{vb})$, is at most of the order of $\log\log n$ [8]. Note that however, even if $E_{X^n}[\min_\theta nH_n(\theta)]$ diverges to minus infinity, this does not necessarily mean $E_{X^n}[nH_n(\overline{\theta}_{vb})]$ diverges in the same order.

Since the dimension of the parameter θ is $MK + K - 1$, the average normalized stochastic complexity of regular statistical models, which coincides with the Bayesian information criterion (BIC) [13] and the minimum description length (MDL) [11], is given by $\lambda_{\mathrm{BIC}}\log n$. Theorem 2 claims that the coefficient $\overline{\lambda}$ of $\log n$ is smaller than λ_{BIC} when α_n is bounded and $\phi_0 \leq 1 + (M-1)/2\alpha^*$. This means that the stochastic complexity $\overline{F}_0(n)$ becomes smaller than the BIC and implies that the advantage of non-regular models in the Bayesian learning still remains in the Variational Bayesian learning.

6. Discussion

In this paper, we showed the upper and lower bounds of the stochastic complexity for mixtures of exponential families in the generalized Variational Bayesian learning.

Firstly let us discuss the lower bound. The lower bound in Eq. (29) can be improved to give

$$\overline{F}_0(n) \geq \overline{\lambda}\log n + E_{X^n}\left[nH_n(\overline{\theta}_{vb})\right] + C_1, \tag{35}$$

if the consistency of the variational estimator $\overline{\theta}_{vb}$ is proven. Note that the coefficient $\overline{\lambda}$ is the same as that of the upper bound given in Theorem 2. Eq. (35) is proved since the consistency implies that $\log \overline{a}_k = O_p(1)$ holds for at least K_0 indexes in Eq. (58). The consistency means that the variational estimator converges to a parameter in the set of the true parameter, $\{\theta | p(x|\theta) = p(x|\theta_0)\}$, with probability 1 as the sample size n is sufficiently large. For some mixture models, the maximum likelihood estimator $\hat{\theta}$ is not consistent [9]. The variational estimator $\overline{\theta}_{vb}$, however, does not necessarily approach the maximum likelihood estimator $\hat{\theta}$ even in the limit $n \to \infty$ and they may have quite different behavior. We conjecture that the variational estimator is consistent and the lower bound in Eq. (35) is obtained for most mixture components. Little has been known so far about the behavior of the variational estimator. Analyzing its behavior and investigating the consistency are important undertakings.

Secondly, in the conventional Bayes case [I] ($\alpha_n = \beta_n = 1$), we compare the stochastic complexity shown in Theorem 2 with the one in the true Bayesian learning. The stochastic complexities in the true Bayesian learning of several non-regular models have been clarified in some recent studies. On the mixture models with M parameters in each component, the following upper bound on the coefficient of the average normalized stochastic complexity $F_0(n)$ in Eq. (14) is known [17,18],

$$\lambda \leq \begin{cases} (K + K_0 - 1)/2 & (M = 1), \\ (K - K_0) + (MK_0 + K_0 - 1)/2 & (M \geq 2), \end{cases} \tag{36}$$

under the same condition (i) about the true distribution and the model described in Section 5 and certain conditions about the prior distribution. Since these conditions about the prior are satisfied by putting $\phi_0 = 1$ in the condition (ii) of Theorem 2, we can compare the stochastic complexity in this case. Putting $\alpha_n = \beta_n = 1$ and $\phi_0 = 1$ in Eq. (31), we have

$$\overline{\lambda} = K - K_0 + (MK_0 + K_0 - 1)/2. \tag{37}$$

Let us compare this $\overline{\lambda}$ of the Variational Bayesian learning to λ in Eq. (36) of the true Bayesian learning. When $M = 1$, that is, each component has one parameter, $\overline{\lambda} \geq \lambda$ holds since $K_0 \leq K$. This means that the more redundant components the model has, the more the Variational Bayesian learning differs from the true Bayesian learning. In this case, $2\overline{\lambda}$ is equal to $2K - 1$ that is the number of the parameters of the model. Hence the BIC [13] and the MDL [11] correspond to $\overline{\lambda} \log n$ when $M = 1$. If $M \geq 2$, the upper bound of λ is equal to $\overline{\lambda}$. Note that $\overline{\lambda}$ and the upper bound of λ are not proportional to M as K grows while λ_{BIC} in Eq. (34) grows proportionally to M. This implies that the variational posterior is close to the true Bayesian posterior when $M \geq 2$. More precise discussion about the accuracy of the approximation can be done for models on which tighter bounds or exact values of the coefficient λ in Eq. (14) are given [14,19].

Thirdly, we point out that Theorem 2 shows how the prior distribution influences the process of the Variational Bayesian learning. By comparing the three cases in Theorem 2, it is obvious that the sequence α_n has much more influence on the result of the learning than the sequence β_n has at least in the case when $\beta_n = o(\log n)$. This implies that one needs to set α_n and β_n with respective appropriate orders instead of setting $\alpha_n = \beta_n$. It is another important issue to assess the dependency of the stochastic complexity on β_n that is proportional to or larger than $\log n$. Moreover, in the case when $\alpha_n = \beta_n = 1$, the coefficient $\overline{\lambda}$ in Eq. (31) is divided into two cases , $\phi_0 \leq (M + 1)/2$ or otherwise, indicating that the influence of the hyperparameter ϕ_0 in the prior $\varphi(\mathbf{a})$ appears depending on the dimension M of the parameter in each component. More specifically, only when $\phi_0 \leq (M + 1)/2$, the prior distribution works to reduce the redundant components that the model has and otherwise it works to use all the components.

And lastly, let us give examples to show how to use the theoretical bounds in the main theorem. Comparing the theoretical bounds in Theorem 2 with experimental results, one can investigate the properties of the actual iterative algorithm in the Variational Bayesian learning. Although the actual iterative algorithm gives the variational posterior that satisfies Eqs. (18) and (19), it may converge to local minima of the functional $\overline{F}[q]$. Remember that Eqs. (18) and (19) are just a necessary condition for $\overline{F}[q]$ to be minimized. One can examine experimentally whether the algorithm converges to the optimal variational posterior that minimizes the functional instead of local minima by comparing the experimental results with the theoretical bounds. The theoretical bounds would also enable us to compare the accuracy of the Variational Bayesian learning with that of the Laplace approximation or the MCMC method. Furthermore, as mentioned in Section 4, the stochastic complexity $\overline{F}(X^n)$ is used as a criterion for the model selection in the Variational Bayesian learning. Our result is important for developing effective model selection methods using $\overline{F}(X^n)$.

7. Conclusion

In this paper, we mathematically proved the lower and upper bounds of the stochastic complexity of the Variational Bayesian learning in mixtures of general exponential families with the generalized prior distribution. These bounds will be used for evaluation and optimization of variational learning systems.

Acknowledgements

This work was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for JSPS Fellows 16-4637 and for Scientific Research 15500130, 2005. A previous version of this paper appeared in the proceedings of the 16th International Conference on Algorithmic Learning Theory, Lecture Notes in Artificial Intelligence, Vol.3734, Springer, Berlin, Heidelberg, 2005.

Appendix. Proof of Theorem 2

From Lemma 1, it is noted that we can evaluate the normalized stochastic complexity $\overline{F}_0(X^n)$ by analyzing two terms $K(r(\theta|\overline{\theta})||\varphi_n(\theta))$ and $(\log C_Q(\overline{\theta}) + S(X^n))$ respectively. First, we evaluate the first one. Since the variational posterior satisfies $r(\theta|\overline{\theta}) = r(\mathbf{a}|\overline{\mathbf{a}})r(\mathbf{b}|\overline{\mathbf{b}})$, we have

$$K(r(\theta|\overline{\theta})||\varphi_n(\theta)) = K(r(\mathbf{a}|\overline{\mathbf{a}})||\varphi_n(\mathbf{a})) + \sum_{k=1}^{K} K(r(b_k|\overline{b}_k)||\varphi_n(b_k)). \tag{38}$$

$K(r(b_k|\overline{b}_k)||\varphi_n(b_k))$ is evaluated as follows.[4]

[4] In this proof, $O_p(1)$ denotes a random variable bounded in probability.

Lemma 2.

$$K(r(b_k|\overline{b}_k)||\varphi_n(b_k)) = \frac{M}{2}\log(n_k+\xi_n) - \log\varphi_n(\overline{b}_k) + O_p(1).$$

Proof of Lemma 2. Using the variational posterior, Eq. (23), we obtain

$$K(r(b_k|\overline{b}_k)||\varphi_n(b_k)) = -\log\frac{C(\gamma_k,\overline{\mu}_k)}{C(\xi_n,\nu_0)} + n_k\{\nu_k\langle b_k\rangle_{r(b_k|\overline{b}_k)} - \langle g(b_k)\rangle_{r(b_k|\overline{b}_k)}\}, \tag{39}$$

where we put $\gamma_k = n_k + \xi_n$. Let us now evaluate the value of $C(\gamma_k,\overline{\mu}_k)$ when γ_k is sufficiently large. From Condition (iii), using the saddle point approximation, we obtain

$$C(\gamma_k,\overline{\mu}_k) = \exp\Big[\gamma_k\{\overline{\mu}_k\cdot\hat{b}_k - g(\hat{b}_k)\}\Big]\sqrt{\frac{2\pi}{\gamma_k}}^{M}\sqrt{|I(\hat{b}_k)|}^{-1}\left\{1+O_p\left(\frac{1}{\gamma_k}\right)\right\}, \tag{40}$$

where $\hat{b}_k$ is the maximizer of the function $\overline{\mu}\cdot b_k - g(b_k)$, that is,

$$\frac{\partial g(\hat{b}_k)}{\partial b_k} = \overline{\mu}_k.$$

Therefore, $-\log C(\gamma_k,\overline{\mu}_k)$ is evaluated as

$$-\log C(\gamma_k,\overline{\mu}_k) = \frac{M}{2}\log\frac{\gamma_k}{2\pi} + \frac{1}{2}\log|I(\hat{b}_k)| - \gamma_k(\overline{\mu}_k\cdot\hat{b}_k - g(\hat{b}_k)) + O_p\left(\frac{1}{\gamma_k}\right). \tag{41}$$

Applying the saddle point approximation to $b_k - \hat{b}_k = \frac{1}{C(\gamma_k,\overline{\mu}_k)}\int(b_k-\hat{b}_k)\exp\{\gamma_k(\overline{\mu}_k\cdot b_k - g(b_k))\}\mathrm{d}b$, we obtain

$$||\overline{b}_k - \hat{b}_k|| \le \frac{C'}{\gamma_k} + O_p\left(\frac{1}{\gamma_k\sqrt{\gamma_k}}\right), \tag{42}$$

where C' is a constant. Since

$$g(b_k) - g(\hat{b}_k) = (b_k-\hat{b}_k)\overline{\mu}_k + \frac{1}{2}(b_k-\hat{b}_k)^{\mathrm{T}}I(b_k^*)(b_k-\hat{b}_k), \tag{43}$$

for some point b_k^* on the line segment between b_k and $\hat{b}_k$, we have

$$g(\overline{b}_k) - g(\hat{b}_k) = (\overline{b}_k-\hat{b}_k)\overline{\mu}_k + O_p\left(\frac{1}{\gamma_k^2}\right), \tag{44}$$

and applying the saddle point approximation we obtain

$$\langle g(b_k)\rangle_{r(b_k|\overline{b}_k)} - g(\hat{b}_k) = (\overline{b}_k-\hat{b}_k)\overline{\mu}_k + \frac{M}{2\gamma_k} + O_p\left(\frac{1}{\gamma_k\sqrt{\gamma_k}}\right). \tag{45}$$

From Eqs. (44) and (45)

$$\langle g(b_k)\rangle_{r(b_k|\overline{b}_k)} - g(\overline{b}_k) = \frac{M}{2\gamma_k} + O_p\left(\frac{1}{\gamma_k\sqrt{\gamma_k}}\right). \tag{46}$$

Thus, from Eqs. (39), (41), (44) and (45), we obtain the lemma. □

Then, the first term on the right-hand side of Eq. (21) is given in the following.

Lemma 3.

$$K(r(\theta|\overline{\theta})||\varphi_n(\theta)) = G(\overline{\mathbf{a}}) - \sum_{k=1}^{K}\log\varphi_n(\overline{b}_k) + O_p(\alpha_n) + O_p(\beta_n) \tag{47}$$

holds where we define the function $G(\overline{\mathbf{a}})$ *of* $\overline{\mathbf{a}} = \{\overline{a}_k\}_{k=1}^K$ *by*

$$G(\overline{\mathbf{a}}) = \frac{MK + K - 1}{2} \log n + \left\{ \frac{M}{2} - \left(\phi_n - \frac{1}{2} \right) \right\} \sum_{k=1}^{K} \log \overline{a}_k. \tag{48}$$

Proof of Lemma 3. From Eqs. (2) and (22), we obtain

$$K(r(\mathbf{a}|\overline{\mathbf{a}})||\varphi_n(\mathbf{a})) = \sum_{k=1}^{K} h(n_k) - n\,\Psi(n + K\phi_n) + \log \Gamma(n + K\phi_n) + \log \frac{\Gamma(\phi_n)^K}{\Gamma(K\phi_n)}, \tag{49}$$

where $\Psi(x) = \Gamma'(x)/\Gamma(x)$ is the di-gamma(psi) function and we used

$$\langle \log a_k \rangle_{r(\mathbf{a}|\overline{\mathbf{a}})} = \Psi(n_k + \phi_n) - \Psi(n + K\phi_n)$$

and the notation $h(x) = x\,\Psi(x + \phi_n) - \log \Gamma(x + \phi_n)$.

By using inequalities for the di-gamma function $\Psi(x)$ and the log-gamma function $\log \Gamma(x)$, for $x > 0$ [2],

$$\frac{1}{2x} < \log x - \Psi(x) < \frac{1}{x}, \tag{50}$$

$$0 \le \log \Gamma(x) - \left(x - \frac{1}{2} \right) \log x + x - \frac{1}{2} \log 2\pi \le \frac{1}{12x}, \tag{51}$$

we obtain

$$h(x) = -\left(\phi_n - \frac{1}{2} \right) \log(x + \phi_n) + x + \phi_n + O(1).$$

Hence, from Eqs. (49) and (51), we obtain

$$\begin{aligned} K(r(\mathbf{a}|\overline{\mathbf{a}})||\varphi_n(\mathbf{a})) \;=\; & -\sum_{k=1}^{K} \left(\phi_n - \frac{1}{2} \right) \log(n_k + \phi_n) + \left(K\phi_n - \frac{1}{2} \right) \log(n + K\phi_n) \\ & + K\phi_n - K\phi_n \log K + \frac{1}{2} \log K + O_p(1), \end{aligned} \tag{52}$$

From Eqs. (38) and (52) and Lemma 2, we complete the proof. □

Let us now turn to the second term, $\log C_Q(\overline{\theta})$, on the right-hand side of Eq. (21). It is evaluated as follows.

Lemma 4.

$$nH_n(\overline{\theta}) + O_p(1) \le -(\log C_Q(\overline{\theta}) + S(X^n)) \le n\overline{H}_n(\overline{\theta}) + O_p(1) \tag{53}$$

holds where the function $H_n(\theta)$ *is defined in Eq.* (11) *and*

$$\overline{H}_n(\overline{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{p(x_i|\theta_0)}{\sum_{k=1}^{K} \overline{a}_k c(x_i|\overline{b}_k) \exp\left\{ -\frac{C}{n_k + \min\{\phi_n, \xi_n\}} \right\}},$$

where C *is a constant.*

Proof of Lemma 4.

$$\begin{aligned} C_Q(\overline{\theta}) \;&= \prod_{i=1}^{n} \sum_{k=1}^{K} \exp\left\langle \log a_k c(x_i|b_k) \right\rangle_{r(\theta|\overline{\theta})} \\ &= \prod_{i=1}^{n} \sum_{k=1}^{K} \exp\{ \Psi(n_k + \phi_n) - \Psi(n + K\phi_n) + \overline{b}_k \cdot f(x_i) - \langle g(b_k) \rangle_{r(\theta|\overline{\theta})} + f_0(x_i) \}. \end{aligned}$$

Using again the inequalities (50) and Eq. (46), we obtain

$$\log C_Q(\overline{\theta}) \geq \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \overline{a}_k c(x_i|\overline{b}_k) \exp\left\{ -\frac{M+2}{2(n_k + \min\{\phi_n, \xi_n\})} + O_p\left(\frac{1}{n_k\sqrt{n_k}}\right)\right\}\right] + O_p(1),$$

$$\log C_Q(\overline{\theta}) \leq \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \overline{a}_k c(x_i|\overline{b}_k) \right] + O_p(1),$$

which give the upper and lower bounds in Eq. (53) respectively. □

From above lemmas, we show the following theorem on the upper bound in Eq. (29).

Theorem 3. *The normalized stochastic complexity $\overline{F}_0(X^n)$ in Eq. (20) satisfies the following inequalities.*
If $\alpha_n \to \alpha^* < \infty$ as $n \to \infty$, then

$$\overline{F}_0(X^n) \leq \overline{\lambda} \log n + O_p(\beta_n),$$

where $\overline{\lambda}$ is given by Eq. (31).
If $\alpha_n \to \infty$, $\alpha_n / \log n \to 0$ as $n \to \infty$, then

$$\overline{F}_0(X^n) \leq \lambda_{\mathrm{BIC}} \log n + O_p(\alpha_n) + O_p(\beta_n),$$

where λ_{BIC} is given by Eq. (34).

Proof of Theorem 3. First, we focus on $\log \varphi_n(\overline{\mathbf{b}}) = \sum_{k=1}^{K} \log \varphi_n(\overline{b}_k)$ in Eq. (47). From the definition of $\varphi_n(b_k)$ Eq. (7),

$$\log \varphi_n(\overline{b}_k) = \beta_n \{\log C(\xi_0, \nu_0) + \log \varphi(\overline{b}_k)\} - \log C(\xi_n, \nu_0). \tag{54}$$

By using the saddle point approximation, as we obtained Eq. (41), $\log C(\xi_n, \nu_0) = O_p(\beta_n)$ follows. We obtain from Eq. (54),

$$\sum_{k=1}^{K} \log \varphi_n(\overline{b}_k) = \beta_n \sum_{k=1}^{K} \log \varphi(\overline{b}_k) + O_p(\beta_n). \tag{55}$$

Then from Lemmas 1, 3 and 4 and Eq. (55), it follows that

$$\overline{F}_0(X^n) \leq \min_{\overline{\theta}} T_n(\overline{\theta}) + O_p(\alpha_n) + O_p(\beta_n), \tag{56}$$

where

$$T_n(\overline{\theta}) = G(\overline{\mathbf{a}}) - \beta_n \sum_{k=1}^{K} \log \varphi(\overline{b}_k) + n\overline{H}_n(\overline{\theta}).$$

From Eq. (56), it is noted that the function values of $T_n(\overline{\theta})$ at specific points of the variational parameter $\overline{\theta}$ give upper bounds of the normalized stochastic complexity $\overline{F}_0(X^n)$. Hence, let us consider the following two cases where R_1 and R_2 are random variables of the order of $\frac{1}{\sqrt{n}}$ such that the constraint $\sum_{k=1}^{K} n_k \nu_k = \sum_{i=1}^{n} f(x_i)$ is met. In both cases, $\sum_{k=1}^{K} \log \varphi(\overline{b}_k) = O_p(1)$ holds.

(I): When

$$\overline{a}_k = a_k^* \quad (1 \leq k \leq K_0 - 1), \quad \overline{a}_k = a_{K_0}^*/(K - K_0 + 1) \quad (K_0 \leq k \leq K),$$

$$\overline{b}_1 = b_1^* + R_1, \quad \overline{b}_k = b_k^* \quad (2 \leq k \leq K_0 - 1), \quad \overline{b}_k = b_{K_0}^* \quad (K_0 \leq k \leq K),$$

then $\overline{H}_n(\overline{\theta}) = O_p(\frac{1}{n})$ holds and

$$T_n(\overline{\theta}) = \frac{MK + K - 1}{2} \log n + O_p(\alpha_n) + O_p(\beta_n).$$

(II): When

$$\overline{a}_k = a_k^* \frac{n + K_0\phi_n}{n + K\phi_n} \quad (1 \le k \le K_0), \quad \overline{a}_k = \frac{\phi_n}{n + K\phi_n} \quad (K_0 + 1 \le k \le K),$$

$$\overline{b}_1 = b_1^* + R_2, \quad \overline{b}_k = b_k^* \quad (2 \le k \le K_0), \quad \overline{b}_k = \nu_0 \quad (K_0 + 1 \le k \le K),$$

then $\overline{H}_n(\overline{\theta}) = O_p(\frac{\alpha_n}{n})$ holds and

$$T_n(\overline{\theta}) = \left\{(K - K_0)\phi_n + \frac{MK_0 + K_0 - 1}{2}\right\} \log n + O_p(\alpha_n) + O_p(\beta_n).$$

The cases (I) and (II) lead to the first inequality of the theorem. The second one follows from the case (I). □

Next we show the following theorem on the lower bound in Eq. (29).

Theorem 4. *The normalized stochastic complexity $\overline{F}_0(X^n)$ in Eq. (20) satisfies the following inequalities.*
If $\alpha_n \to \alpha^* < \infty$ as $n \to \infty$, then

$$\overline{F}_0(X^n) \ge \underline{\lambda} \log n + nH_n(\overline{\theta}_{vb}) + O_p(\beta_n), \tag{57}$$

where $\underline{\lambda}$ is given by Eq. (30).
If $\alpha_n \to \infty$, $\alpha_n / \log n \to 0$ as $n \to \infty$, then

$$\overline{F}_0(X^n) \ge \lambda_{\mathrm{BIC}} \log n + nH_n(\overline{\theta}_{vb}) + O_p(\alpha_n) + O_p(\beta_n),$$

where λ_{BIC} is given by Eq. (34).

Proof of Theorem 4. Since $\log \varphi(\overline{b}_k)$ is bounded, it follows from Lemmas 1, 3 and 4 and Eq. (55),

$$\overline{F}_0(X^n) \ge \min_{\overline{\mathbf{a}}}\{G(\overline{\mathbf{a}})\} + nH_n(\overline{\theta}_{vb}) + O_p(\alpha_n) + O_p(\beta_n). \tag{58}$$

If $\phi_n > \frac{M+1}{2}$, then

$$G(\overline{\mathbf{a}}) \ge \frac{MK + K - 1}{2} \log n - \left(\frac{M+1}{2} - \phi_n\right) K \log K, \tag{59}$$

since Jensen's inequality yields that $\sum_{k=1}^K \log \overline{a}_k \le K \log(\frac{1}{K})$.
If $\phi_n \le \frac{M+1}{2}$, then

$$G(\overline{\mathbf{a}}) \ge \left\{(K-1)\phi_n + \frac{M}{2}\right\} \log n + O_p(\alpha_n), \tag{60}$$

since $\overline{a}_k \ge \frac{\phi_n}{n + K\phi_n}$ holds for every k and the constraint $\sum_{k=1}^K \overline{a}_k = 1$ ensures that $\log \overline{a}_k = O_p(1)$ for at least one index k. From Eqs. (58)–(60), we obtain the theorem.

Let us combine these theorems and complete the proof. □

Proof of Theorem 2. From Theorems 3 and 4, if α_n is bounded, we have

$$\underline{\lambda} \log n + nH_n(\overline{\theta}_{vb}) + O_p(\beta_n) \le \overline{F}_0(X^n) \le \overline{\lambda} \log n + O_p(\beta_n),$$

and if $\alpha_n \to \infty$, $\alpha_n / \log n \to 0$ $(n \to \infty)$, then

$$\lambda_{\mathrm{BIC}} \log n + nH_n(\overline{\theta}_{vb}) + o_p(\log n) \le \overline{F}_0(X^n) \le \lambda_{\mathrm{BIC}} \log n + o_p(\log n).$$

Taking expectations over all sets of training samples, we obtain Theorem 2. □

References

[1] H. Akaike, Likelihood and Bayes procedure, in: J.M. Bernald (Ed.), Bayesian Statistics, University Press, Valencia, Spain, 1980, pp. 143–166.
[2] H. Alzer, On some inequalities for the Gamma and Psi functions, Mathematics of Computation 66 (217) (1997) 373–389.
[3] H. Attias, Inferring parameters and structure of latent variable models by variational Bayes, in: Proceedings of Uncertainty in Artificial Intelligence (UAI'99), 1999.
[4] M.J. Beal, Variational algorithms for approximate Bayesian inference, Ph.D. Thesis, University College London, 2003.
[5] L.D. Brown, Fundamentals of statistical exponential families, in: IMS Lecture Notes-Monograph Series, vol. 9, 1986.
[6] B.S. Clarke, A.R. Barron, Information-theoretic asymptotics of Bayesian methods, IEEE Transactions on Information Theory IT-36 (3) (1990) 453–471.
[7] Z. Ghahramani, M.J. Beal, Graphical models and variational methods, in: D. Saad, M. Opper (Eds.), Advanced Mean Field Methods — Theory and Practice, MIT Press, 2000.
[8] J.A. Hartigan, A failure of likelihood asymptotics for normal mixtures, in: Proceedings of the Berkeley Conference in Honor of J. Neyman and J. Kiefer, vol. 2, 1985, pp. 807–810.
[9] J. Kiefer, J. Wolfowitz, Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters, Annals of Mathematical Statistics 27 (4) (1956) 887–906.
[10] D.J. Mackay, Bayesian interpolation, Neural Computation 4 (2) (1992) 415–447.
[11] J. Rissanen, Stochastic complexity and modeling, Annals of Statistics 14 (3) (1986) 1080–1100.
[12] M. Sato, Online model selection based on the variational Bayes, Neural Computation 13 (7) (2001) 1649–1681.
[13] G. Schwarz, Estimating the dimension of a model, Annals of Statistics 6 (2) (1978) 461–464.
[14] K. Watanabe, S. Watanabe, Stochastic complexities of Gaussian mixtures in variational Bayesian approximation, Journal of Machine Learning Research 7 (2006) 625–644.
[15] K. Watanabe, S. Watanabe, Stochastic complexities of general mixture models in variational Bayesian learning, International Journal of Neural Networks 20 (2) (2007) 210–219.
[16] S. Watanabe, Algebraic analysis for non-identifiable learning machines, Neural Computation 13 (4) (2001) 899–933.
[17] K. Yamazaki, S. Watanabe, Singularities in mixture models and upper bounds of stochastic complexity, International Journal of Neural Networks 16 (2003) 1029–1038.
[18] K. Yamazaki, S. Watanabe, Stochastic complexity of Bayesian networks, in: Proceedings of Uncertainty in Artificial Intelligence (UAI'03), 2003.
[19] K. Yamazaki, S. Watanabe, Newton diagram and stochastic complexity in mixture of binomial distributions, in: Proceedings of Algorithmic Learning Theory (ALT2004), 2004, pp. 350–364.